Chapitre 5 : Choix de modèle

On rappelle que, dans le modèle linéaire, un **individu** de la population est donc représenté par le vecteur $(X_1, \dots, X_p, Y)$. On suppose que

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

où $\varepsilon \perp\!\!\!\perp (X_1, \dots, X_p)$, $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$. Les paramètres de ce modèle sont $\beta_0, \dots, \beta_p, \sigma^2$.

Malgré son apparante simplicité, le modèle linéair a des avantages en termes d'**interprétabilité** et souvent il fournit de bonnes **performances prédictives**. L'ajustement de ces modèles se fait par moindres carrés ou par maximum de vraisemblance.

Nous avons déjà vu deux alternatives à l'ajustement par moindres carrés :

- **ridge**, pour traiter les cas où les covariables sont fortement corrélées ($\iff$ matrice de corrélation a des valeurs propres proches de 0)
- **lasso**, pour supprimer les covariables les moins importantes.

Ici, nous allons chercher à supprimer des covariables de la liste, pour trouver un meilleur modèle. Nous allons distinguer deux cas ici :

— **interprétation** : on cherche le modèle qui ne contient que les covariables nécessaires. Dans ce cas, les variables exclues n'apportent aucune information complémentaires sur Y au delà des variables que l'on conserve.
— **prédiction** : on cherche le modèle qui a les meilleures qualités prédictives sur de nouvelles données.

1 Critères de sélection

On peut utiliser différents critères pour comparer des modèles. Pour l'instant, on peut imaginer que l'on compare deux modèles $M_{(1)}$ et $M_{(2)}$ construits à partir de deux sous-ensembles des covariables. Voici quelques critères.

— R^2 **ajusté** : on cherche à viser le meilleur modèle en termes de prédiction. Mais, même corrigé, le critère de R^2 reste trop optimiste quant aux qualités prédictives d'un modèle. Ce n'est pas un bon critère, sauf s'il y a peu de covariables, et beaucoup de données.
— **BIC** : *Bayesian Information Criterion* ou critère de Schwartz. Il s'agit d'un critère qui visible à trouver le « vrai » modèle qui a engendré les données,

et doit donc être utiliser si on veut obtenir un modèle **interprétable**.

— **AIC** : *Akaike Information Criterion*. Il s'agit d'un critère qui construit un modèle avec de bonnes **qualités prédictives**, en réalisant un compromis entre
  — le **biais**, éventuellement introduit en otant des covariables qui auraient pu être utiles et
  — la **variance des coordonnées** de $\widehat{\boldsymbol{\beta}}$ lorsque le nombre de covariables est grand.

— **erreur de généralisation calculée par validation croisée** : ce genre de critères visent également des **qualités prédictives**, mais sont parfois un peu lent en temps de calcul.

Sur le modèle $M_{(i)}$, notons
— $\mathbf{X}_{(i)}$ la matrice de design du modèle,
— $\boldsymbol{\beta}_{(i)}, \sigma^2_{(i)}$ les paramètres du modèle,
— $\widehat{\boldsymbol{\beta}}_{(i)}, \widehat{\sigma}^2_{(i)}$ leurs estimations par maximum de vraisemblance.

Les critères BIC et AIC sont définis en partant de la valeur de la log-vraisemblance maximale. Rappelons que la **log-vraisemblance** est donnée par

$$\ell_{(i)}(\boldsymbol{\beta}_{(i)}, \sigma^2_{(i)}) = -\frac{1}{2\sigma^2_{(i)}} \left\| \mathbf{Y} - \mathbf{X}_{(i)} \boldsymbol{\beta}_{(i)} \right\|^2 - \frac{n}{2} \log \sigma^2_{(i)}.$$

et son maximum vaut

$$\widehat{\ell}_{(i)} = -\frac{1}{2\widehat{\sigma}^2_{(i)}} \left\| \mathbf{Y} - \mathbf{X}_{(i)} \widehat{\boldsymbol{\beta}}_{(i)} \right\|^2 - \frac{n}{2} \log \widehat{\sigma}^2_{(i)}.$$

L'estimateur du maximum de vraisemblance est également l'estimateur des moindres carrés. Ainsi, si le modèle $M_{(1)}$ est strictement inclus dans le modèle $M_{(2)}$, on a

$$\widehat{\ell}_{(2)} > \widehat{\ell}_{(1)}$$

On ne peut donc pas choisir i en maximisant la log-vraisemblance $\widehat{\ell}_{(i)}$.

On définit les **critères de vraisemblance pénalisés** par

$$\begin{aligned} \mathrm{AIC}_{(i)} &= 2d_{(i)} - 2\widehat{\ell}_{(i)}, \\ \mathrm{BIC}_{(i)} &= d_{(i)} \log n - 2\widehat{\ell}_{(i)}, \end{aligned}$$

où $d_{(i)} = \mathbf{rg}(\mathbf{X}_{(i)})$. (Attention, le signe de la définition change d'un livre ou d'un logiciel à l'autre).

Le meilleur modèle au sens de ces critères est celui qui maximise leurs valeurs.

Remarque. Dans le cadre de la régression linéaire, le critère des « C_p de Mallow » est équivalent au critère AIC :

$$C_p = \frac{1}{n}\Big(SSE_{(i)} + 2d_{(i)}\widehat{\sigma}^2_{(i)}\Big).$$

Le critère du R^2 **ajusté** est défini par

$$R^2 \text{ ajusté} = 1 - \frac{SSE/(n-d_{(i)}-1)}{SST/(n-1)}.$$

2 Algorithmes de sélections de variables

Dans toute cette partie, on suppose le critère fixé.

2.1 Parcours exhaustif

L'objectif de cet algorithme est de comparer tous les modèles possibles à partir des covariables $X_1, \ldots, X_p$.

1. Notons M_0 le modèle nul, qui ne contient que l'intercept. Ce modèle prédit Y avec $\bar{Y}$.
2. Pour chaque valeur de k entre 1 et p :
   (a) Ajuster chacun des $\binom{p}{k}$ modèles linéaires à k covariables,
   (b) Choisir parmi ces modèles le meilleur modèle au sens du plus petit SSE ou du plus grand R^2, et le nommer M_k
3. Parmi les modèles $M_0, M_1, \ldots, M_p$ choisir le meilleur modèle au sens du critère choisi.

Cette méthode ne convient que si le nombre total p de covariables est faible.

— Elle est coûteuse car elle nécessite d'ajuster de très nombreux modèles.

— Plus le nombre de modèles ajustés est grand, plus les chances de trouver un modèle qui semble correct uniquement sur les données, avec de mauvaises propriétés de généralisation est grand. Cela peut donc conduire à un problème de **sur-apprentissage**.

Pour ces deux raisons, on utilise plutôt des méthodes pas à pas qui ne regardent pas tous les sous-modèles du modèles complets.

2.2 Parcours pas-à-pas

Sélection progressive L'idée est de partir du modèle nul M_0 qui ne contient que l'intercept, et d'ajouter progressivement des covariables.

1. Notons M_0 le modèle nul, qui ne contient que l'intercept. Ce modèle prédit Y avec $\bar{Y}$.
2. Pour chaque valeur de k entre 0 et $p-1$:
   (a) Ajuster tous les $p-k$ modèles linéaires qui consistent à ajouter une nouvelle covariables à M_k.
   (b) Choisir parmi ces modèles le meilleur modèle au sens du plus petit SSE ou du plus grand R^2, et le nommer M_{k+1}
3. Parmi les modèles $M_0, M_1, \ldots, M_p$ choisir le meilleur modèle au sens du critère choisi.

Sélection rétrograde L'idée est de partir du modèle complet M_p qui contient toutes les covariables et d'enlever progressivement des covariables.

1. Notons M_p le modèle complet, qui contient toutes les covarialbes
2. Pour chaque valeur de k entre p et 1 :
   (a) Ajuster tous les k modèles linéaires qui consistent à enlever une covariables à M_k.
   (b) Choisir parmi ces modèles le meilleur modèle au sens du plus petit SSE ou du plus grand R^2, et le nommer M_{k-1}
3. Parmi les modèles $M_0, M_1, \ldots, M_p$ choisir le meilleur modèle au sens du critère choisi.

3 Réduction de dimension

On peut également utiliser des méthodes de réduction de dimension sur la matrice $\mathbf{X}$ pour réduire le nombre de covariables.

L'ACP permet de se concentrer sur quelques variables (les premiers axes). On parle alors de régression sur composantes principales. Mais ces axes sont obtenus par une méthode **non supervisée** puisque la réponse Y n'influe pas sur le calcul

de ces axes. Il n'y a aucune garantie que les axes principaux soient les meilleures variables pour prédire la réponse Y.

La méthode **Partial Least Square** (PLS) est une méthode de réduction de la dimension qui construit de nouvelles variables Z_1, Z_2,... qui sont des combinaisons linéaires des variables originales comme dans l'ACP. On suppose ici que les covariables X_j sont centrées-réduites.

Voici une idée du fonctionnement de PLS. La première variable est

$$Z_1 = \sum_{i=1}^{p} \varphi_{1,j} X_j$$

où $\varphi_{1,j}$ est l'effet estimé lorsque l'on régresse Y sur la seule covariable X_j.

Les directions suivantes sont obtenues en prenant les résidus de la régression de Y sur Z_1 et en répétant la règle ci-dessus en utilisant ces résidus comme nouvelle réponse à prédire...

Dans la plupart des cas, PLS est la méthode la plus efficace pour trouver un modèle ayant de bonnes qualités prédictives.